Policy Sciences 16 (1983) 147–163
Elsevier Sience Publishers, B.V., Amsterdam – Printed in The Netherlands

On Evaluating Evaluations*

RICHARD C. LARSON

Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, U.S.A.

LENI BERLINER

Miraud Associates, Inc., Bethesda, MD, U.S.A.

ABSTRACT

The act of evaluation requires an expenditure of resources. In Part I of this paper, we present a simple decision tree model borrowed from operations research to provide a conceptual framework for considering whether or not to commit such resources. In Part II, once the evaluation is carried out, we address the problem of evaluating the evaluation as a vehicle for producing useful information to decisionmakers. Evaluation inputs, processes, and outcomes are defined and discussed within the context of comprehensive evaluation of evaluations.

I. Introduction

Because the stakes are so high with public programs, both with regard to monetary expenditures and to achieving various social goals, evaluation of public programs (as an aid to implementation) has assumed an evermore prominent role during the past, say, fifteen years. Using techniques and paradigms borrowed from diverse substantive and methodological areas of concern, "evaluation" has emerged as a somewhat strange amalgam. Noticeably lacking are unifying theories, constructs and paradigms derived for and descriptive of evaluation itself. Our goal in this paper is to offer two simple constructs related to evaluating an evaluation. Both constructs are motivated by our decision-oriented interpretation of evaluation, derived from the following logic:

* This work was supported by grant 80-IJ-CX-0048 from the National Institute of Justice of the U.S. Department of Justice to the Operations Research Center of the Massachusetts Institute of Technology. Points of view or opinions stated in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

1. Evaluation is a process that produces information;
2. Information is useful only to the extent that it informs decisions;
3. A decision is an irrevocable allocation of resources [1];
4. Thus, evaluation is a process that produces information to assist in the allocation of resources.

In the first construct we place the decision – "evaluate" or "do not evaluate" – within a very simple decision tree. Several intrinsic properties of evaluations are elucidated even by this most simple of examples. And, even though the example is borrowed (from operations research) and simple, we show that it is structurally a generalization of the popular "two-alternative-hypothesis" evaluation design so often seen in the evaluation literature. While our first construct is conceptual (not operational) in nature and focuses on a priori analysis of one or more proposed evaluation designs, our second construct deals with realistic issues to be addressed in evaluating conducted evaluations. Our concern here is motivated in large part by our observation that those who critique evaluations tend to do so with a narrow technical focus: "Was the t-test carried out properly? Are the statistical procedures designed in such a way so as to be biased in favor of proving the null hypothesis?" Our message here too is simple: just as programs are to be evaluated comprehensively – involving an integrated analysis of program inputs, processes and outcomes – so too evaluations should be evaluated comprehensively – involving analysis of *evaluation* inputs, processes and outcomes. We suggest an illustrative set of evaluation inputs, processes and outcomes; reflecting the decision point of view, the outcomes are decisions influenced by information provided by the evaluation.

II. A Decision Maker: Buyer of a Used Car

The decision-influencing role of an evaluator can be simply demonstrated by recourse to a favorite problem of operations research involving the buyer of a used car. We describe this problem and its solution not because it incorporates all the subtleties and ambiguities of an actual evaluative setting, but because it does not: it contains a very simplified abstraction of several basic elements of the evaluation process; by studying the abstraction, we can gain insight for more complex situations.

We consider an individual faced with a resource allocation problem. She wants to purchase a used car and simultaneously wants to minimize expected costs. She decides to purchase her car from Avertz Car Rent, which sells its fleet of one-year old cars each September. Upon entering the Avertz lot she gazes upon a virtually limitless supply of used cars, each costing $4,000. While these cars all appear identical to our prospective car buyer, previous buyers have discovered that one third are in fact *lemons*; each lemon will require an additional $2,000 of repair costs. The remaining two-thirds are *peaches* requiring no repair costs. To our happy (hapless?) car buyer, lemons and peaches are indistinguishable.

Our car buyer knows of an *evaluator* of used cars who, for \$200, will examine a car and pronounce it either a peach or a lemon. But like all evaluators, this evaluator is imperfect: he correctly identifies 90 percent of lemons and 80 percent of peaches. (These facts regarding the evaluator's performance were made available free of charge by the Independent Association of Evaluators of Evaluators).

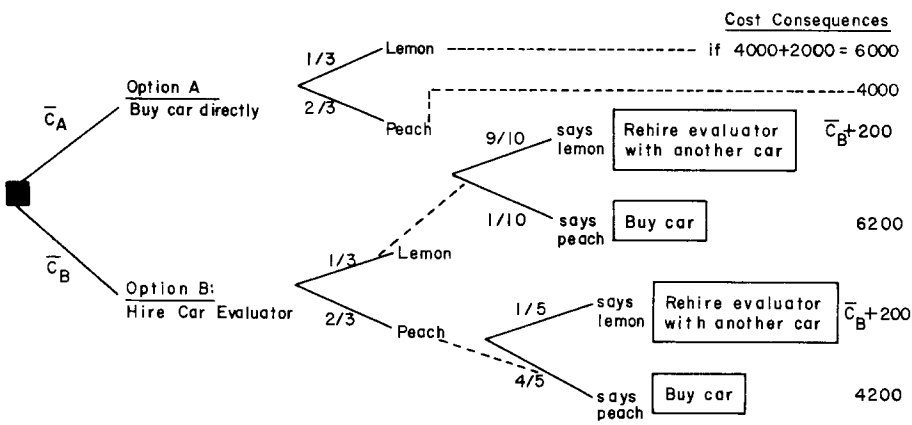
If the evaluator is employed and if he says "lemon" for a given car, our car buyer decides not to buy that car but to select another one and have that one evaluated at an additional cost of \$200. This process continues until the evaluator first says "peach," at which time our car buyer buys that particular car.

Here is the decision question: *Should the car buyer hire the evaluator?*

Decision Tree Analysis

The answer to the question above can be determined by analyzing the decision tree shown in Fig. 1. The car buyer has two options: Option A, buy the car directly; or Option B, hire the evaluator. Under Option A, there are two possible consequences: with a probability of 1/3, the purchased car is a lemon, requiring a total of \$4000 (purchase) plus \$2000 (for repairs); with a probably of 2/3, the purchased car is a peach, requiring a total cost of only \$4000. The expected cost of Option A is thus

$$\bar{C}_A = \frac{1}{3}(6000) + \frac{2}{3}(4000) = \$\,4667.$$



$$\bar{C}_A = \frac{1}{3}(6000) + \frac{2}{3}(4200) = \$4667$$

$$\bar{C}_B = \frac{1}{3}\left[\frac{9}{10}(\bar{C}_B + 200) + \frac{1}{10}(6200)\right] + \frac{2}{3}\left[\frac{1}{5}(\bar{C}_B + 200) + \frac{4}{5}(4200)\right]$$

$$\therefore \bar{C}_B = \$4470$$

Fig. 1. Decision tree analysis for used car buyer.

Under Option B, there is still a 1/3 chance that any car examined will be a lemon and a 2/3 chance that it will be a peach. Suppose first that the car is a lemon. If the evaluator errs and pronounces the car a peach, then the car buyer must pay \$4000 (purchase) plus \$200 (evaluation fee) plus \$2000 (repairs) = \$6200. If on the other hand the evaluator correctly identifies the lemon, then the car buyer avoids buying that defective car, but at a cost of \$200 for the evaluative information. Moreover, the car buyer must go back to the car lot and select another car for testing, in effect restarting "from ground zero" with Option B. If $\bar{C}_B$ is the expected total cost associated with Option B (total cost including car purchase, repair, and evaluator's fee), then the car buyer when she hears "lemon" pronounced associates with that pronouncement \$200 (for the immediate fee) plus an expected additional future cost of $\bar{C}_B$ (for going back to ground zero). Similar logic yields, for the case in which the examined car is a peach, a cost of \$4200 when the evaluator says "peach" and $\$200 + \bar{C}_B$ when the evaluator says "lemon." Combining all these costs with the appropriate probabilities, one obtains a linear equation for the unknown $\bar{C}_B$,

$$\bar{C}_B = \frac{1}{3}\left[\frac{9}{10}(\bar{C}_B + 200) + \frac{1}{10}(6200)\right] + \frac{2}{3}\left[\frac{1}{5}(\bar{C}_B + 200) + \frac{4}{5}(4200)\right]$$

Solving this equation, we obtain

$$\bar{C}_B = \$4470.$$

If the car buyer is willing to compare the two alternatives on a basis of expected monetary cost, then Option B ("hire the evaluator") saves the car buyer an expected amount equal to \$4667–4470 = \$197. Thus, a "rational decisionmaker," using expected monetary value as a decision criterion, would have an evaluation performed to reduce uncertainty and risk about the ultimate decision, in this case purchasing a car.

One can continue to analyze the decision tree under a number of different assumptions to gain further insight. For instance, if the evaluator were perfect (i.e., never made errors) yet charged the same fee, then the expected cost of Option B would be

$$\bar{C}_B' = \frac{1}{3}(\bar{C}_B' + 200) + \frac{2}{3}(4200),$$

implying

$$\bar{C}_B' = \$4300.$$

In this case the potential expected savings when compared to Option A is $\bar{C}_A - \bar{C}_B' = \$367$. Since the expected savings with the imperfect evaluator is only \$197, we may conclude that (\$367–\$197) = \$170 is the amount of potential savings lost due to imperfections of the originally described evaluator. One could continue the analysis by asking the following question: "What is the maximum that the car buyer should

ever be willing to pay the evaluator?" Clearly, when the cost of Option B ($\bar{C}_B$) exceeds the cost of Option A ($\bar{C}_A$), then evaluation is no longer the preferred option. By setting $\bar{C}_B$ equal to $\bar{C}_A$ and letting the evaluator's fee be an unknown quantity, say x, we can determine the maximum amount we should be willing to pay. For the case of the perfect evaluator, this computation is

$$\bar{C}_B{}' = \bar{C}_A = \$4667 = \frac{1}{3}(4667 + x) + \frac{2}{3}(4000 + x);$$

solving for x, we obtain $x = \$455$ as the maximum fee that we should ever be willing to pay the perfect evaluator. A similar analysis for the imperfect evaluator yields a maximum fee of \$318. In general, the amount one is willing to pay for an evaluation decreases as the quality of the information produced by the evaluation decreases.

Looking back at the analysis we see that the information provided free of charge about the evaluator's performance was of value to the decisionmaker; one could recast the entire decision tree allowing for a payment to obtain that information (i.e., paying for information which is essentially an evaluation of the evaluator).

If our car buyer had selected Option B with the imperfect evaluator and if she had been "unlucky," with pronouncements of three "lemons" followed by "peach," when in fact the last car turned out to be a lemon, the final cost would have been $4 \times \$200 =$ \$800 (evaluator's fees) + \$4000 (purchase) + \$2000 (repair) = \$6800, much greater than $\bar{C}_A = \$4667$. Retrospectively, someone evaluating her actions might say that she selected unwisely. But any such "Monday morning quarterbacking" must be done from the perspective of the decisionmaker's information profile at the time the decision was made; from this point of view, she still made the correct decision (given expected monetary value as the decision criterion). The key points of this exercise can be summarized as follows:

1. The decision to evaluate is itself an allocation of resources which can only be justified if the expected benefits outweigh the expected costs of the evaluation.
2. Imperfection of an evaluator reduces the expected benefit of the evaluation.
3. However, even imperfect information (if not "too imperfect") is better than no information.
4. In certain instances, it is a rational allocation of resources to spend money to evaluate an evaluator.
5. Retrospective analysis of a decision must be based on the decision maker's information profile at the time of the decision, not on information that subsequently became known.

Comparison with a Classical Paradigm

While the car buyer example may seem a bit far fetched and oversimplified, it actually contains the structure inherent in every classical two-alternative-hypothesis evalua-

tion. Suppose a governmental agency is deciding whether or not to implement a particular program. If the program is implemented, only two outcomes are possible:

H_0: the program has no effect;
H_1: the program has a beneficial effect.*

From all relevant information, the governmental agency assesses the likelihood of H_0 to be P_0; the likelihood of H_1 is $P_1 = 1 - P_0$. The agency has the option of either implementing the program directly (and enduring "nature's coin flip" according to the probabilities P_0 and P_1) or commissioning an evaluation, the outcome of which will be one of two statements:

Statement 0: "The program is likely to have no effect,"

or

Statement 1: "The program is likely to have a beneficial effect."

If the evaluator is hired and eventually says Statement 0, then the agency decides *not* to implement the program; if the evaluator says Statement 1, then the agency goes ahead with the program implementation.

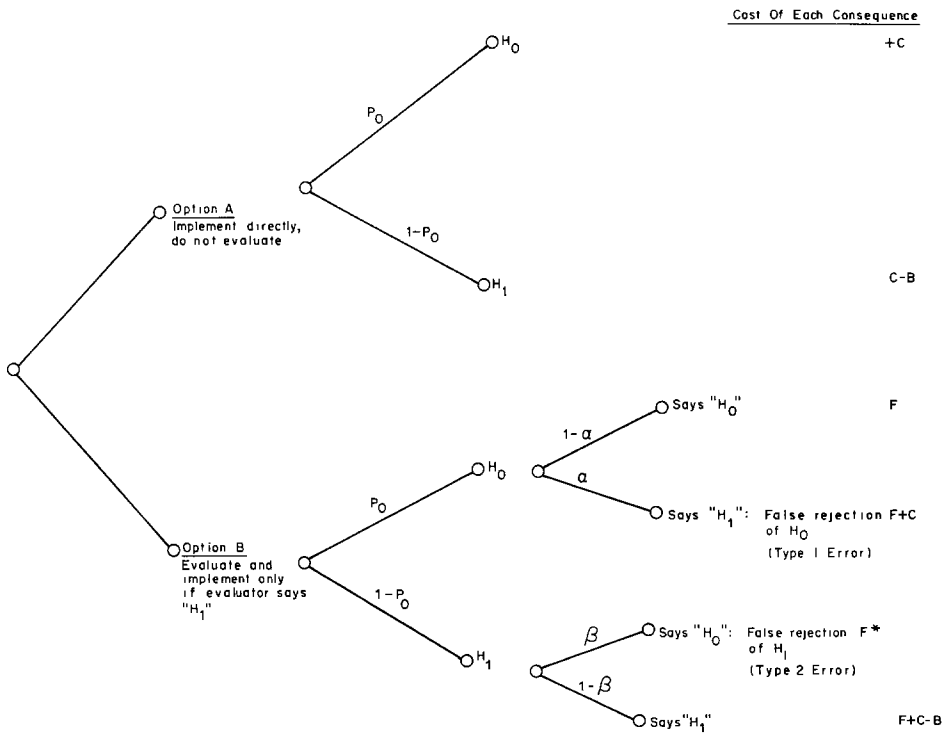


Fig. 2. Decision tree analysis for two-alternative hypothesis evaluation.

* For simplicity of presentation, we are ignoring negative effects of the program.

Like any evaluator, the hired evaluator is imperfect. If in fact H_0 is the true state of affairs, there is a probability α that the evaluator errs and says that the program is likely to have an effect. If in fact H_1 is the true state of affairs, there is a probability β that the evaluator errs and says that the program is likely to have no effect.

The decision structure above is depicted in Fig. 2. It is identical to that of the used car buyer example, except that it is simpler! There is no mechanism here for "repeatedly selecting new and different programs," analogous to repeated sampling of cars.

We can complete this structuring of the two-hypothesis evaluation by imposing an appropriate cost structure. Suppose we define costs as follows:

C = cost of implementing the program (in dollars).
B = societal benefit of the program, expressed in dollars, given that H_1 is the true state of affairs.*
F = fee of the evaluator.

Then the costs associated with each of the six alternative outcomes shown in Fig. 2 are shown in the right-most column in Fig. 2. For example, $C - B$ is the net program cost (implementation cost minus societal benefits) if H_1 is true and no evaluator is employed; presumably $B > C$, so that $C - B < 0$. $F + C - B$ is the analogous cost for the situation in which an evaluator is hired and correctly deduces that H_1 is true. If, on the other hand, the evaluator falsely rejects H_1, there is a fee of F incurred plus an "opportunity loss" of $B - C$ (due to the fact that society will not gain its net benefit of $B - C$ which is possible by implementing the program).

Analyzing the two options, the expected cost of option A is

$$C_A = C - (1 - P_0)\, B$$

The expected cost of option B, ignoring opportunity loss, is

$$C_B = F + C\,[P_0\,\alpha + (1 - P_0)\,(1 - \beta)] - \beta\,(1 - P_0)\,(1 - \beta)$$

Here, if $C_B < C_A$, then it makes sense to hire the evaluator who has conditional error probabilities of α and β for H_0 and H_1, respectively. To determine the maximum fee (F_{max}) that we should ever pay the evaluator, we solve for $F = F_{max}$ when setting $C_B = C_A$. The result is

$$F_{max} = C_A - \{C[P_0\,\alpha + (1 - P_0)\,(1 - \beta)] - B\,(1 - P_0)\,(1 - \beta)\}$$

We have thus shown one can analyze the two-hypothesis evaluation that one so often sees in the literature by utilizing a simple decision tree structure, incorporating Bayesian probability estimates, costs of alternative outcomes (expressed in commen-

* It is assumed that the societal benefit is 0 if the program is implemented and H_0 is true.

surate terms), and the evaluator's performance characteristics and fee. Here, of course, α is the conditional probability of a "Type 1 error" and represents the "level of significance" of the evaluation test(s) when in fact H_0 is true; β is the conditional probability of a "Type 2 error." Much of the criticism of simple two-hypothesis evaluation structures can be placed in this framework, including biasing of an experimental design in favor of "proving the null hypothesis," not recognizing the costs of evaluation errors, and not including the cost of the evaluation.

In actual evaluation settings, life is almost always more complicated than that situation depicted above. Costs and benefits of a program are difficult to estimate accurately prior to implementation and rarely can be expressed in commensurate terms. Often there are more than two possible "states of nature." The evaluator may offer a range of fees, each associated with a different α and β. Or, the evaluator's findings may be constructive, leading to an improved program design and thus to one or more successive stages of the decision tree. Even with such complications, the insights provided, first by the numerical "used car buyer" example, and second by the generic Bayesian hypothesis testing example, are valuable regarding the decision to evaluate or not to evaluate.

Why Evaluate Evaluations?

Evaluations are performed to provide information about a program to decisionmakers. Evaluations of evaluations (EOE's) are performed to provide information about an evaluation to a possible different set of decisionmakers. For instance, an EOE can provide an independent assessment for decisionmakers of the quality of the information presented in the evaluation. This would enhance the extent of "informedness" of the resulting decisions, for a cost – the cost of the EOE. Clearly if this cost exceeds some threshold, its marginal information value may not be adequate to justify its expenses. In this context, the decision to perform an EOE is also an allocation of resources which may or may not be justified at a particular point in time, given one's knowledge about the original evaluation, the program being evaluated, in terms of the marginal cost and expected marginal information content of the EOE.

While considerable attention has been given to the evaluation of programs, it is somewhat ironic that relatively little attention has been devoted to evaluating the evaluations (and the evaluators). Exceptions are Bernstein and Freeman (1975); Cook and Gruder (1978); the U.S. General Accounting Office (1978a, b); Minnesota Systems Research Inc. (1973); Cook (1978); and Stufflebeam (1974). Many, perhaps most, analyses of evaluations have focussed on one sub-element of the evaluation, namely the technical aspects of evaluation process, as reflected, say, by the statistical methodology employed or the survey research methods used. This limited focus creates problems analogous to those which would occur if one conducted an evaluation of the program, exclusively examining some part of program process, while ignoring program input, outcomes and other aspects of process. In the same sense that

evaluation of programs requires the comprehensive analysis of program inputs, process, and outcome, so too evaluative study of evaluations requires analysis of evaluation inputs, process, and outcome. In this section, we lay out the elements of such an approach.

Three Components of Evaluation

Any evaluation is a process, having inputs and yielding outcomes. Any comprehensive EOE should examine all three of these evaluation components. An examination of process alone may, for instance, verify exemplary technique, but reveals nothing about evaluation impact. A review of evaluation outcomes alone is not sufficient for explaining the causal mechanisms linking evaluation input through process to those outcomes; indeed, one who analyzes only outcomes is often hardpressed to attribute outcomes to the evaluation. An examination of inputs alone reveals little more than the collection of resources mustered to conduct the evaluation. We present below an initial listing of detailed elements comprising evaluation input, process, and outcome. Many of the items discussed under each heading have direct analogies in terms of the used car buyer problem; we leave it as a reader option whether to develop or to ignore these analogies.

Evaluation inputs may be considered to be *an inventory of resources and methodologies brought to bear on the evaluation, and the basic elements of the evaluation/program setting.* One proposed set of inputs is summarized in Table 1, and includes evaluation budget (both in absolute terms and as a percentage of program budget), duration of the evaluation, timing of the evaluation with respect to the program being evaluated, and skills (and other attributes) of the evaluation personnel. Despite the indisputable importance of these items and despite the urgings of evaluators to

TABLE 1

Proposed Inputs to an Evaluation

Evaluation inputs:	An inventory of resources and methodologies brought to bear on the evaluation and the basic elements of the evaluation/program setting.

1. Budget (and other material resources available to the evaluators).
2. Duration
3. Timing with respect to the program being evaluated.
4. Attributes of evaluation personnel (e.g., training, experience, "world view").
5. Attributes of program personnel (e.g., experience, commitment, education).
6. Program attributes (e.g., goals, substantive area of concern, client group).
7. Evaluation methodology and design.
8. Audience, or "client group," and purposes of the evaluation.
9. Existing data and data limitations.
10. Underlying theoretical model(s).
11. The expected policy consequences.

consider program inputs during evaluations, few evaluators themselves document these rudimentary evaluation inputs. Any comprehensive EOE is thwarted at an early stage if these inputs are not known.

Other necessary inputs also listed in Table 1 include attributes of the program being evaluated and its personnel (e.g., training and experience, determining their attitudes toward the evaluation), evaluation methodology and design, audience or client group for the evaluation, and the programmatic purpose of the evaluation. Evaluation methodology and design should describe not only the statistical procedures to be used to analyze data, but also the entire plan for considering program inputs, process and outcomes. Moreover, it should especially indicate where and how information generated by evaluation activities is to be fed back to program staff for possible program modification; rules for such adaptive change, to the extent possible, should be stated explicitly a priori (i.e., such rules for adaptation are themselves evaluation inputs).

The potential value of the evaluation will of course be affected by the quality and quantity of existing data, the cost of collecting additional data, and the possible evaluator/program interaction (and program contamination) created by certain data collection activities.

A primary input to the evaluation would be the aggregate social science knowledge that pertains to the program being evaluated expressed in the form of one or more causal models linking program inputs through process to desired outcome. It is such causal models that both guide the evaluators' activities and provide a theoretical structure against which to compare program operations.

A final primary input to an evaluation is the expected extent of policy improvement to be obtained by conducting the evaluation. In the language of the used car buyer or the two-hypothesis evaluation example, this quantity is $\bar{C}_B - \bar{C}_A$. If $\bar{C}_B < \bar{C}_A$, then the evaluation is expected to more than justify its costs by providing useful decision-oriented policy information leading towards an improved allocation of agency or societal resources. If $\bar{C}_B \geq \bar{C}_A$, then the evaluation is not justified (given the validity of the decision tree analysis). In a more general sense, the quantity we are interested in is the "Prior Information Value" (PIV) of the evaluation. The PIV of an evaluation is the expected policy improvement consequence of the evaluation design, minus the expected costs of implementing the design (Thompson, 1975). Here the policy improvement and the cost units must be compatible, a requirement which should not be taken lightly in the complex field of evaluation. The mathematical operation of calculating expectation by averaging over all alternative outcomes must be done before the evaluation is undertaken, in a fashion similar to that demonstrated by the two simple decision tree examples. This expectation operation requires inclusion of subjective probabilities over outcomes as well as the policy utility of alternative outcomes.

At this stage of evaluation research, the PIV as an input to an evaluation may be limited to a useful conceptual device. The estimation of the expected policy benefits of an evaluation and standardization of the benefits and costs is simply too difficult a task for all but the simplest situations. This limitation of the PIV concept does not,

TABLE 2

Process Components of an Evaluation

Evaluation process:	Actual conduct of the evaluation compared with that planned in the evaluation design.

1. Types, intensity and frequency of interactions between evaluators and program staff members.
2. Response of program staff and client groups to the presence of evaluators.
3. Extent to which acquired information is fed back to program staff, perhaps modifying program procedures.
4. Extent to which acquired information is used to modify the allocation of evaluation resources.
5. Adaptiveness of evaluation design (i.e., capacity to respond to changes in the program).
6. Changes in personnel (e.g., evaluators, program staff, client groups of both program and evaluation).
7. Methodology: the formal and informal processing of information leading to evaluative findings.
8. Communication of findings.

however, preclude its conceptional utility as a means of addressing two questions: 1) whether to evaluate and 2) how to select the appropriate evaluation design.

The above list of evaluation inputs is illustrative; additional inputs could readily be added. The key point, however, is that – analogous to a program – each evaluation is characterized by a set of inputs which a priori can provide significant information to all concerned parties regarding the potential utility of the evaluation. Upon examination, it would not be inappropriate in some instances either to require adjustment in inputs prior to implementing the evaluation or – if that is infeasible – to reverse the decision to conduct the evaluation. Examination of evaluation inputs may thus help explain disappointing or successful evaluation impact and/or process.

Utilization of evaluation inputs is *evaluation process, the actual conduct of the evaluation as compared with that planned in the evaluation design.* One proposed set of components of evaluation process is given in Table 2. Again, this list is meant to be illustrative, not exhaustive. The importance of the first three items is self-explanatory.

The next two items relate to the adaptability of the evaluation. First, one is interested in (4) the extent to which information acquired during the evaluation is used to modify the allocation of evaluation resources. Is the evaluation in a "straitjacket" design, or can the evaluators modify the design in response to information obtained during the course of the evaluation? Adaptability may be reflected in elements of process evaluation such as the allocation of participant observers and/or interviewers to various parts of the program. Or it could relate to the sequential adaptive generation and testing of alternative hypotheses regarding program operation. Many evaluations in practice are adaptive, but lacking rules and encouragement for adaptability, the evaluators in their reports are not eager to describe this element of their evaluation. A related issue is (5), the adaptiveness of the evaluation design in response to changes in the program being evaluated. For instance, during the operation of the program, an employee strike could occur, a new relevant law could be enacted, or a citizens' group

could protest against some particular aspect of the program. To what extent is the evaluation jeopardized by such program changes and interruptions, and to what extent can it adapt to them? No evaluation design can stand impervious to all conceivable unforeseen changes in the program and its operating environment, but some are more robust than others. A chronological history of adaptions of the evaluation to changes in the program would seem to be an important part of evaluating evaluation process.

The next item is (6), changes in personnel (e.g., evaluators, program staff, client groups of both the program and the evaluation). This is one of perhaps several internal unplanned changes in program or evaluation process. However, it is a critical one, in that a turnover in one or more key persons in the evaluation or in the program can markedly affect the outcomes of both. Any evaluation having two or more directors in succession is vulnerable to breaks in continuity of plan and purpose; to an evaluator of such an evaluation, neglect of such a leadership change could lead to erroneous conclusions regarding causes for observed limited evaluation impact. A change in the client group of either the program or the evaluation is also important. For instance, a significant fraction of evaluations that have had little or no eventual decision impact appear to fall victim to the "vanishing advocate" syndrome, in which the person who originally commissioned the evaluation has moved to another professional position, only to be replaced by someone unsympathetic to the original purposes of the evaluation (cf. Chaiken et al., 1975).

The seventh entry in Table 2, (7) methodology, is a pivotal component of evaluation process. It appears to be this component, evaluation technique and methodology as applied in practice, that has received most scrutiny by evaluators of evaluations. Perhaps this is because manipulation of numbers is one of the few elements of evaluation process that can be replicated and scrutinized by others after termination of the evaluation. And statistical procedure appears to be associated with apparently universal "scientific" measures of accountability. But one should not fall prey to the trap of misplaced emphasis on statistical method. A statistically-elegant evaluation may be seriously flawed in other respects and statistical correctness by no means guarantees decision impact. On the other hand, a statistically flawed evaluation may indeed present imperfect information to decisionmakers; the imperfections may lead to decisions that would have been improved if more accurate information had been available; the "costs" of such imperfect information can be considerable. Yet, when balanced with other components of evaluation process, it is quite possible that a statistically-flawed evaluation can still present useful information to decisionmakers – where usefulness implies decisions being made that are in some sense "better" than those that would have been made in the total absence of the evaluation (e.g., the used car buyer). To place in perspective the importance of statistical procedure and to estimate the cost of statistical error, we would argue strongly for a comprehensive evaluation of evaluation process, as reflected by the other elements in Table 2.

The eighth and final entry in the Table is (8) communication of final evaluation

findings. Included here is the final report, its structure, content, level and style of presentation being important parts of the communication process. But also important are oral presentations, use of teaching aids to convey the essential results, and other activities and devices for communication and dissemination of results. For instance, a methodologically flawless evaluation whose findings are unintelligible to its sponsors or clients will have at best marginal impact.

To summarize our discussion of evaluation process, considerable evaluation and program activity evolves over time, much of it unforeseeable at the evaluation design stage. These activities influence the content and quality of evaluative information that is collected, in turn influencing the evaluative findings. Because of this causal chain in the dynamics of an evaluation itself, elements of evaluation process should be made known to potential "consumers" of the evaluation's findings, thereby providing a type of self-reported quality measure to be attached to the findings. Scrutiny of methodology alone is insufficient, representing a narrow focus on only one aspect of evaluation process.

Any attempt to demarcate the boundary between evaluation process and evaluation outcome is done in the presence of ambiguity and controversy. Still, the inherent difficulties should not act to preclude discussion on this vital matter. To provide one input to the debate, we take a firm stand on evaluation outcome, motivated by our decision orientation: *Evaluation outcomes are the decisions (resource allocations) influenced by the evaluation.* Most evaluators discover the decision consequences of their evaluation only long after submission of the final report, if at all; because of this, it is inappropriate for those who evaluate evaluations to judge their effect only from reading the final report (Larson et al., 1979). The time period of the evaluation of an evaluation must extend beyond that of the original evaluation to attempt to assess its ultimate decision consequences.

In Table 3 we have summarized five distinct types of decisions that may be influenced by an evaluation. For each, there are difficult questions relating to (1) the influence of this particular evaluation on the decisionmaker's action versus the role of knowledge gained elsewhere; (2) the fact that any retrospective analysis has a cut-off time, and actions occurring after that time cannot be included in the assessment of decision impact; (3) retention of the status quo is itself a decision, but one particularly difficult to attribute to knowledge gained from an evaluation.

"The funding agency's decision" (to fund, refund, modify or cancel the program) is an obvious one, especially since program funding agencies fund so many evaluations for this very purpose. And problems of attribution and causality are not really so troublesome here as they are with other decisions listed in Table 3. That is, it is not a rare event for a funding agency to refund, modify or cancel a program based on information from an evaluation of that program.

Likewise, "the program's staff operational decision" (to modify any of the program procedures) can in many cases be linked directly with an evaluation, particularly when elements of evaluation process are known.

TABLE 3

Evaluation Outcomes

Evaluation Outcomes:	The decisions influenced by the evaluation.
1.	Decision by funding agency to fund, refund, modify or cancel program.
2.	Decision by program staff to modify any of the program procedures.
3.	Decision by members of the client group to alter participation patterns in the program.
4.	Decision by one or more members of the research community to study further the questions/issues raised in the evaluation.
5.	Decision by one or more other funders and/or program personnel (in other jurisdictions) to initiate, modify, or terminate similar programs.

More difficult to link causally to the evaluation is "the program client group's decision" to alter participation patterns in the program. Evaluations are usually funded by agencies or groups other than program clients, and thus in many (if not most) instances the evaluation findings may not even be available to the clients; if available, the findings may not be widely known throughout the client population. And even if they are known, it is not clear how that knowledge would influence participation patterns. For instance, clients of a human services program that received a primarily negative evaluation may still have no choice about where to receive that service, and thus they would continue their participation even with knowledge of program flaws.

Perhaps the final two decision types listed in Table 3: "the research community's decision" and "the decision of those involved in related programs" have the longest-term impact. Each of these types of decisions appears often to be made on the basis of collective knowledge, any one evaluation contributing perhaps only marginally to the knowledge pool. (See, for example, the literature on "metaevaluation" research, e.g., Glass, 1976 and Rosenthal, 1978.) Researchers and agency administrators in other jurisdictions must at best be called "second-order decisionmakers," since they are not directly involved with the program being evaluated or the evaluation. But given our decision framework, to exclude them would preclude any finding of long term or widespread consequences of an evaluation. As an example, perhaps the largest ultimate impact of the 1936 *Literary Digest Poll*, which predicted Alf Landon to win over Franklin Delano Roosevelt in a landslide, was a commitment by the evaluation and survey research methodology communities to learn more about the threat of selection bias. (The Salk Polio Vaccine trials in the 1950s also had this effect.)

Evaluation Documentation Requirements

The framework just described has implications for documentation of program evaluations. Given our definition of comprehensive evaluation, the information required to

do a comprehensive evaluation of evaluations can become considerable. Whether or not an independent EOE is to be performed, documentation in the final report of inputs, processes and outcomes (as far as possible) of the evaluation is essential to increase its utility to decisionmakers.

As we discovered in an empirical study of criminal justice program evaluations (Larson et al., 1979), current evaluation documentation practice is uneven or sorely lacking. Of the roughly 200 studies in the sample, only 4 percent indicated the percentage of the program budget allocated for the evaluation, and only 2 percent indicated evaluation budget. Thirty-one percent of the reports in the sample reported the total duration of the evaluation, while 8 percent at least indicated (though not always explicitly) the timing of the evaluation with respect to the program being evaluated. None of the reports described the professional or other attributes of either program staff or evaluation personnel. Finally, while 90 percent of the reports made at least some reference to the context or purpose of the effort, actual potential users of the evaluations were rarely identified explicitly. Only 58 percent of the reports contained an analysis of program goals, and only 47 percent discussed the program's client group in any way.

Evaluation process components fared much worse than evaluation inputs in the sample of final reports. The constituent parts of evaluation process listed in Table 2 were rarely if ever included in the final reports. In our review of the evaluation research literature, we have found that elements of evaluation process do appear in the growing number of anecdotal reports on non-utilization (e.g., Weiss, 1977). But there appears to be little tradition of evaluators routinely reporting on their own evaluation process. Such lack of self-reporting reduces the ability of decisionmakers to assess the quality of information produced by the evaluation. Information on evaluation process could only serve to enhance evaluators' and program managers' awareness of evaluation limitations and pitfalls, thus leading to improvement of evaluation practice.

Evaluation outcomes in terms of decisions influenced by the evaluation are rarely documented in the final report, due in part to the timing of the final report with respect to decisions yet to be made. And even after decisions are made, it is often exceedingly difficult to estimate what influence (if any) the evaluation had on the decisions. Here, it seems we need new methods for information feedback during the evaluation, follow-up, attribution and documentation.

Thus, we can at present only make a plea for more complete self-reporting of evaluation inputs and process. Self-reporting is open to criticism on grounds of objectivity, particularly in the area of evaluation process. Yet even imperfect information in this area would be more valuable than the present state of nearly no information. Particularly for those second-order decisionmakers not directly affiliated with the program being evaluated, it seems that at least rudimentary knowledge of evaluation inputs and process would be necessary to assess the possible relevance of the findings to them.

III. Summary and Suggested Research

We have proposed that evaluation is a process producing information that can be evaluated on the basis of its relevance and anticipated benefit to decisionmakers; in times of fiscal constraint, there can be no other justification for program evaluation.

We have proposed two simple constructs that we believe may be of assistance in the evaluating of evaluations. In the first, we utilize a simple decision tree approach to demonstrate certain properties of the situation confronting a decisionmaker who is anticipating paying for an evaluation. This example demonstrated that the decision to evaluate is itself an allocation of resources which can only be justified if the expected benefits outweigh the expected costs of the evaluation. It also showed how imperfect evaluations reduce the anticipated benefits for the decisionmaker, but that even imperfect information is often useful in carrying out improved decisionmaking. It can also be a rational allocation of resources to spend money to evaluate an evaluator. And, when looking at past decisions to conduct evaluations, one must attempt to replicate the state of information available to the decisionmaker at the time of the evaluation decision. This first simple construct also was shown to be a generalization of the classic two-alternative hypothesis evaluation design. Issues of costs of evaluation error, cost of evaluator, and biasing of the evaluation design in various ways, can all be addressed within such a simple decision analytic framework. But also as discussed in Section I, most often evaluations are much more complicated than those illustrated by the simple decision trees in Figs. 1 and 2, thereby relegating these constructs more to the realm of conceptualization than to implementation.

Paralleling program operation, an evaluation can be characterized by inputs, process, and outcomes. In our second construct, we provide lists of each, arguing that the only ultimate outcomes of an evaluation are decisions influenced by the evaluation. Thus, however difficult to measure, the effect of an evaluation must be judged on the basis of resources (re)allocated as a consequence of evaluation information provided. Evaluations of evaluations per se have several potential purposes: to provide an independent assessment to decisionmakers of the quality of information contained in an evaluation; to provide guidance in selecting an evaluator; to assimilate "research knowledge" from a number of separate but similar programs; and to provide a vehicle for examining the evaluation enterprise itself. Our concern for evaluation inputs, processes, and outcomes extends to recommendations for improved evaluation documentation in these areas.

Further work is needed to devise methods for carrying out comprehensive evaluations of evaluations, within time and budget constraints that are acceptable to potential decisonmakers. Numerous important questions abound: Who should conduct evaluations of evaluations? When is self-reporting of evaluation inputs and process adequate? How do we measure the effect of evaluation information on a decision? How do we historically recreate a decisionmaker's state of (imperfect) knowledge at the time of decision? Should different evaluation criteria be applied to evaluations that

were performed after program implementation and to those done during program implementation? In the latter case, is it fair to expect greater "decision-impact"? How much greater? If a follow-on program is instituted to measure an evaluation's ultimate impact, who should fund it and who should do it? Each of these questions provide fruitful areas for future study.

Notes

1 This definition of decision is taken from Howard (1966). "'Irrevocable' does not imply 'for all time,' but at least for the next short time interval that an allocation of at least one resource has been made. That is, a decision is not a 'decision to make a decision,' but rather the concrete action implied by the decision. After any time interval, a decision may be replaced by another decision, perhaps based on updated information."

2 The small Roman numeral in parentheses identifies the particular point in Table 2.

References

Bernstein, I. N. and Freeman, H. E. (1975). *Academic and Entrepreneurial Research: The Consequences of Diversity in Federal Evaluation Studies.* New York: Russell Sage Foundation.

Chaiken, J., Cranbill, L., Holliday, L., Jaquette, D., Lawless, M. and Quade, E. (1975). "Criminal Justice Models: An Overview." Santa Monica, CA: Rand Corporation Technical Report R-1859-DOJ.

Cook, T. D. (1978) "Utilization, knowledge-building, and institutionalization: Three criteria by which evaluation research can be evaluated," (Introduction) in Cook, T. D. (ed.), *Evaluation Studies Review Annual*, Vol. 3. Beverly Hills: Sage.

Cook, T. D. and Gruder, C. L. (1978). "Metaevaluation research," *Evaluation Quarterly* 2: 5–51.

Howard, R. A. (1966). "Decision Analysis: Applied Decision Theory," in Hertz, D. B. and Melese, J. (eds.) *Proceedings of the Fourth International Conference on Operations Research.* New York: Wiley.

Glass, G. U. (1976). "Primary, secondary and meta-analysis of research," *Educational Researcher*, 5 (10): 3–8.

Larson, R. C., Bier, V. M., Kaplan, E. H., Mattingly, C., Eckels, T. J., Rechman, N. and Berliner, L. S. (1979). "Interim Analysis of 200 Evaluations on Criminal Justice." (LEAA Grant 78NI-AX-0007). Operations Research Center. Cambridge, MA.: Massachusetts Institute of Technology.

Larson, R. C. and Kaplan, E. H. (1981). "Decision oriented approaches to program evaluation," *New Directions for Program Evaluation*, 10: 49–68.

Minnesota Systems Research, Inc. (1973). "Proposed Review Systems for the Evaluation of Technical Quality & R & D Projects in the Department of Health Education and Welfare."

Rosenthal, R. (1978) "Combining Results of Independent Studies," *Psychological Bulletin* 85, No. 1, 185–193.

Rossi, P. H. and Wright, S. R. (1977) "An assessment of theory, practice, and politics," *Evaluation Research* 1: 5–52.

Stufflebeam, D. L. (1974). "Meta-Evaluation." Michigan University Evaluation Center. Occasional Paper #3.

The Literary Digest, Vol. 121, No. 1, January 4, (1936). Funk & Wagnalls Co., 354–360 Fourth Avenue, New York.

Thompson, M. (1975). *Evaluation for Decisions in Social Programmes.* Westmead, England: Saxon House, D.C. Heath Ltd.

Thompson, M. (1982). *Decision Analysis For Program Evaluation.* Cambridge, MA: Ballinger Publishing Co.

United States General Accounting Office (1978a) "Status and Issues: Federal Program Evaluation." (PAD-78-83). Washington, D.C.

United States General Accounting Office (1978b). "Assessing Social Program Impact Evaluations: A Checklist Approach." (Exposure Draft) (PAD-79-2), Washington, D.C.

Weiss, C. H. (1977). *Using Social Research in Public Policy Making.* Lexington, MA.: Lexington Books.